

Numériser le patrimoine I: standards et bonnes pratiques

Les Bonnes pratiques

Simon Gabay

Genève, 22 septembre 2020

Introduction

.* virtuelles

Les données patrimoniales concernent notamment

- Le texte
- L'image fixe
- L'image animée
- Le son

Il serait impossible de tout couvrir dans un seul cours, nous nous proposons donc de nous intéresser au patrimoine littéraire ancien en tant qu'objet (manuscrit, imprimé...).

Les enjeux posés par la numérisation des fonds patrimoniaux des bibliothèques reste néanmoins très similaires à ceux d'autres institutions similaires. On parle ainsi de *GLAM*, (*galleries, libraries, archives, museums*).

Au delà de l'interface

Dans la pratique, la confrontation avec le patrimoine numérisé consiste essentiellement en la consultation de sites. Est-ce agréable à consulter? Voire joli? Est-ce pratique à utiliser? Trouve-t-on l'information facilement? Pourrait-on proposer des services aux utilisateurs?

Les spécialistes des humanités numériques et les informaticiens s'intéressent beaucoup à l'utilisation des applications en ligne qu'ils développent. On parle ainsi d'*interface*, d'*interaction homme-machine* ou d'*expérience utilisateur*.

Au delà de ces questions évidentes (et importantes), il existe un autre problème plus technique: celui de l'intéropérabilité des données, de leur standardisation, de la pérennisation... C'est un versant plus technique sur lequel nous aimerions nous pencher dans ce cours.

Prenons un exemple, celui d'[e-codices](#).

Cologny, Fondation Martin Bodmer, Cod. Bodmer 40



Vielliard Françoise, Manuscrits Français du Moyen Âge, Cologny-Genève, 1975, pp. 32-34.

[Aperçu](#)[Fac-similé](#)[Afficher PDF](#)[Afficher XML](#)[Imprimer description](#)**Titre du manuscrit:** Jean Bodel, Chanson des Saisnes**Période:** Fin du XIII^e siècle.**Support:** Parchemin**Volume:** II + 122 + II feuillets.**Format:** 176 × 124 mm.**Numérotation des pages:** Foliotation récente, partie à l'encre, partie au crayon.**Composition des cahiers:** Cahiers: 1-10⁸, 11⁹ [répartis 4-5], 12-14⁸, 15⁹ [répartis 5-4] sans réclames.**Etat:** Les deux feuillets de garde, [au début](#) et [à la fin](#), sont contemporains de la reliure.**Mise en page:** Justification env. 140 × 75 mm. Texte sur deux colonnes; 32 lignes par colonne. Réglerie à la mine de plomb.**Type d'écritures et copistes:** Ecriture gothique. Une seule main.**Décoration:** Initiales de laisse (2 lignes) rouges; initiales à la plume rouges et noires à filigranes et antennes des deux couleurs:

- au fol. [1](#) (9 lignes),
- au fol. [21v^o](#) (5 lignes),

Cologny, Fondation Martin Bodmer, Cod. Bodmer 40



Vielliard Françoise, Manuscrits Français du Moyen Âge, Cologny-Genève, 1975, pp. 32-34.

[Aperçu](#)[Fac-similé](#)[Afficher PDF](#)[Afficher XML](#)[Imprimer description](#)**Titre du manuscrit:** Jean Bodel, Chanson des Saisnes**Période:** Fin du XIII^e siècle.**Support:** Parchemin**Volume:** II + 122 + II feuillets.**Format:** 176 × 124 mm.**Numérotation des pages:** Foliotation récente, partie à l'encre, partie au crayon.**Composition des cahiers:** Cahiers: 1-10⁸, 11⁹ [répartis 4-5], 12-14⁸, 15⁹ [répartis 5-4] sans réclames.**Etat:** Les deux feuillets de garde, [au début](#) et [à la fin](#), sont contemporains de la reliure.**Mise en page:** Justification env. 140 × 75 mm. Texte sur deux colonnes; 32 lignes par colonne. Réglerie à la mine de plomb.**Type d'écritures et copistes:** Ecriture gothique. Une seule main.**Décoration:** Initiales de laisse (2 lignes) rouges; initiales à la plume rouges et noires à filigranes et antennes des deux couleurs:

- au fol. [1](#) (9 lignes),
- au fol. [21v^o](#) (5 lignes),

Under the hood

```
<sourceDesc>
  <bibl>PDF vorhanden</bibl>
  <msDesc xml:lang="fra" xml:id="fmb-cb-0040">
    <msIdentifier>
      <settlement>Cologny</settlement>
      <repository>Fondation Martin Bodmer</repository>
      <idno>Cod. Bodmer 40</idno>
    </msIdentifier>
    <head>
      <title>
        <persName role="author" key="pnd_118660454">Jean Bodel</persName>
        , Chanson des Saisnes
      </title>
      <origDate notBefore="1270" notAfter="1299">Fin du XIIIe siècle</origDate>
    </head>
    <msContents>
      <msItem>
        <locus from="1r" to="122v">Ff. 1–122v°</locus>
        <author key="pnd_118660454">Jean Bodel</author>
        <title>Chanson des Saisnes</title>
      </msItem>
    </msContents>
  </msDesc>
</sourceDesc>
```

Bonnes pratiques

Numériser n'est pas compliqué, le faire dans les règles de l'art l'est beaucoup plus. Il s'agit donc de comprendre la raison de ces règles et d'apprendre à les appliquer.

Ces "bonnes pratiques" sont compliquées à suivre, car elles impliquent de suivre des règles génériques, et donc qui ne sont pas spécifiquement adaptées à nos problèmes. C'est un effort supplémentaire dans un premier temps, mais cela simplifie considérablement le travail par la suite: c'est cela, les humanités numériques.

Points de repère

Les standards

Il est fondamentale de suivre, autant que possible, des standards. Par exemple:

- ISO (*International Organization for Standardization*): Organisation internationale de normalisation
- Unicode (ISO/CEI 10646) pour le codage de texte écrit
- DC (*Dublin core*, ou ISO_15836) est un socle commun d'éléments descriptifs (auteur, date, lieu...)
- TEI (*Text Encoding Initiative*): recommandations pour l'encodage de documents textuels. Il existe un équivalent pour la musique (MEI).

Les métadonnées

Les métadonnées accompagnent les données qu'elles enrichissent, parfois dans un fichier apparenté à la donnée (image), parfois à l'intérieur même des données (transcription)

- Les métadonnées *techniques* sont souvent générées automatiquement par l'appareil qui produit la donnée (paramètres de prise de vue, réglage de l'appareil...)
- Les métadonnées *descriptives* ciblent le contenu du document. Elles décrivent le contenu (auteur, titre, date...)

Exemple des métadonnées *Dublin core* au format XML d'un portrait photographique de Zora Neale Hurston (l'original se trouve [ici](#)).

```
<!DOCTYPE dublinCore PUBLIC '-//OCLC//DTD Dublin core v.1.
<dublinCore>
  <title>[Portrait of Zora Neale Hurston]</title>
  <author type='photographer'>Van Vechten, Carl</author>
  <otherAgent type='digitizer'>Any Library</otherAgent>
  <subject scheme='gmgpc'>Portrait Photographs</subject>
  <objectType>image</objectType>
  <form scheme='IMT'>image/jpeg</form>
  <relation type='ammemParent'>vanv</relation>
  <identifier type='URN'>hdl:loc.pp.vanv/5a52142</identif.
</dublinCore>
```

Les formats

Une donnée numérique est enregistrée dans un format qui implique une déformation plus ou moins grande, et donc un poids plus ou moins important. Les formats diffèrent selon les types de données:

- Texte: `.xml` , `.doc` , `.docx` ...
- Image: `.tiff` , `.jpg` , `.png` ...
- Video: `.avi` , `.mp4` , `.mpeg` ...
- Audio: `.mp3` , `.wav` ...

Il existe des formats propriétaires (`.doc`) et des formats ouverts (`.docx`), des formats dépréciés et des nouveaux...

Partager

De plus en plus on tente de construire des ponts entre les projets. On distingue ainsi le lieu où les données sont stockées et celui où l'on peut la trouver, voire la réutiliser:

- OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*)
- IIIF (*International Image Interoperability Framework*) pour l'interopérabilité pour la diffusion et l'échange d'images haute résolution sur le Web
- DTS (*Distributed Text Services*) pour l'interopérabilité pour la diffusion et l'échange de transcriptions sur le Web

Si l'on reprend notre manuscrit de la Bodmer, il est possible de récupérer toutes les images avec le lien suivant:

<https://www.e-codices.unifr.ch/metadata/iiif/fmb-cb-0040/i>

En allant sur un autre site que celui de la Bodmer, par exemple celui de la BNF (<https://demos.biblissima.fr/mirador/>), on peut afficher les images suisses, qui restent stockées à Genève.

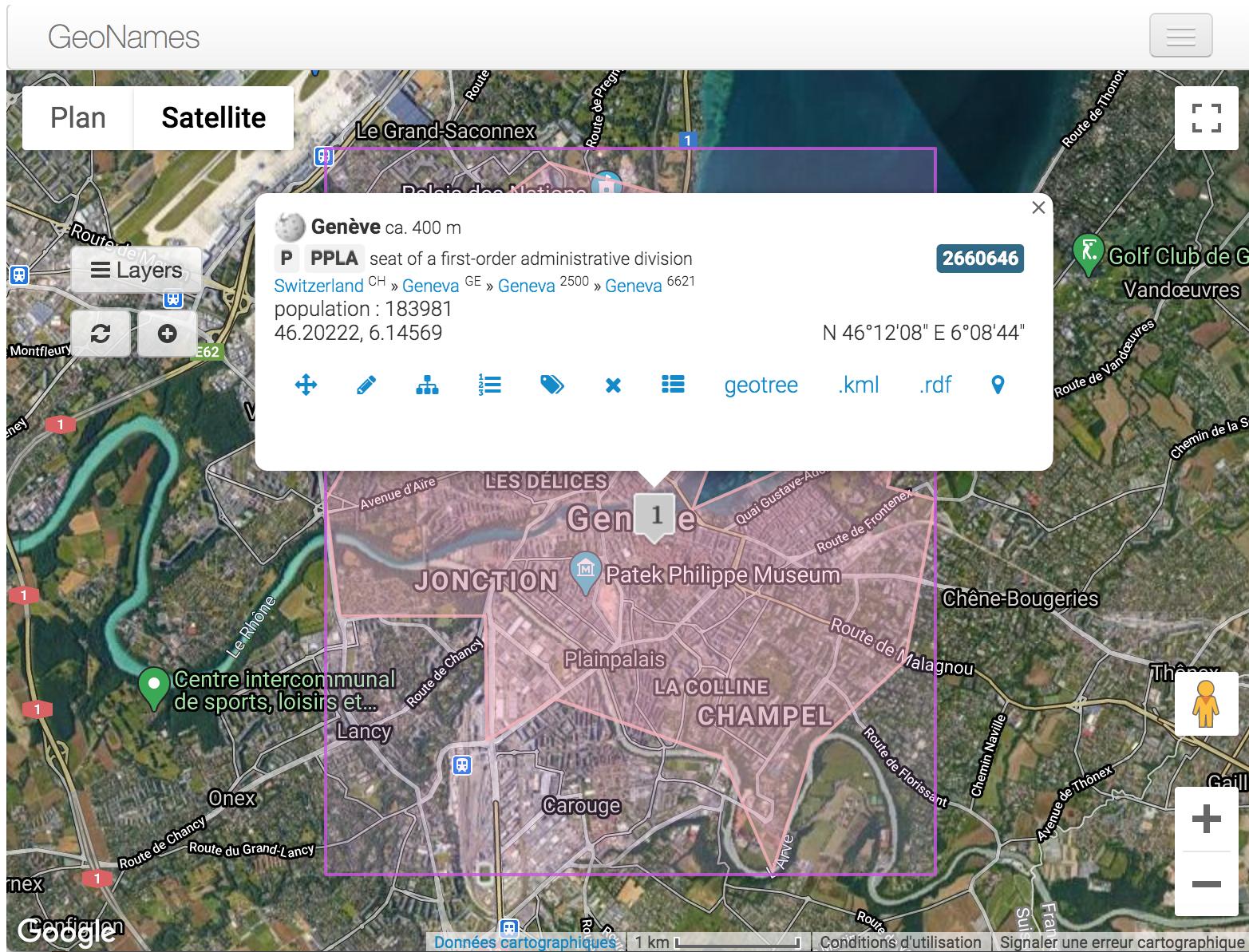
```
"canvases": [
  {
    "@id": "https://www.e-codices.unifr.ch/metadata#I00000000000000000000000000000000",
    "@type": "sc:Canvas",
    "label": "Front cover",
    "height": 6496,
    "width": 4872,
    "images": [
      {
        "@id": "https://www.e-codices.unifr.ch/metadata#I00000000000000000000000000000000",
        "@type": "oa:Annotation",
        "motivation": "sc:painting",
        "on": "https://www.e-codices.unifr.ch/metadata#I00000000000000000000000000000000",
        "resource": {
          "@id": "https://www.e-codices.unifr.ch/metadata#I00000000000000000000000000000000",
          "@type": "dctypes:Image",
          "format": "image/jpeg",
          "height": 6496,
          "width": 4872,
          "service": {
            "@context": "http://iiif.io/api/image",
            "@id": "https://www.e-codices.unifr.ch/metadata#I00000000000000000000000000000000",
            "profile": "http://iiif.io/api/image/2.0/context.json"
          }
        }
      }
    ]
  }
]
```

Les référentiels

Les entités uniques, comme les personnes, les lieux, les objets... peuvent avoir des identifiants uniques permettant de faciliter la réutilisation et la connection des données entre les projets:

- ISNI (*International Standard Name Identifier*) pour les personnes
- GeoNames pour les noms de lieux
- ISMI (*International Standard Manuscript Identifier*) pour les manuscrits, sur le modèle de l'ISBN

Fiche GeoNames de Genève (<https://www.geonames.org/2660646>)



Les droits

On utilise des licences qui permettent de protéger son travail, et de respecter celui des autres. Est-ce que je peux:

- Vendre des données en ligne?
- Modifier les données que j'ai trouvées?
- Diffuser des données sans indiquer la source?
- Diffuser ces données sous une autre forme que je les ai trouvés?

Il existe plusieurs solutions:

- Creative commons
- Etalab
- MIT
- ...
-

Exemple de licence sur Wikipedia:

Conditions d'utilisation

Objet

Ceci est une reproduction photographique fidèle d'une œuvre d'art originale en deux dimensions. L'œuvre d'art elle-même est dans le [domaine public](#) pour la raison suivante :



L'auteur est mort en 1519 ; cette œuvre est donc également dans le [domaine public](#) dans tous les pays pour lesquels le [droit d'auteur a une durée de vie de 100 ans ou moins après la mort de l'auteur](#).

Cette œuvre est dans le [domaine public](#) aux États-Unis car elle a été publiée avant le 1^{er} janvier 1925.

Ce fichier a été identifié comme étant exempt de restrictions connues liées au droit d'auteur, y compris tous les droits connexes et voisins.

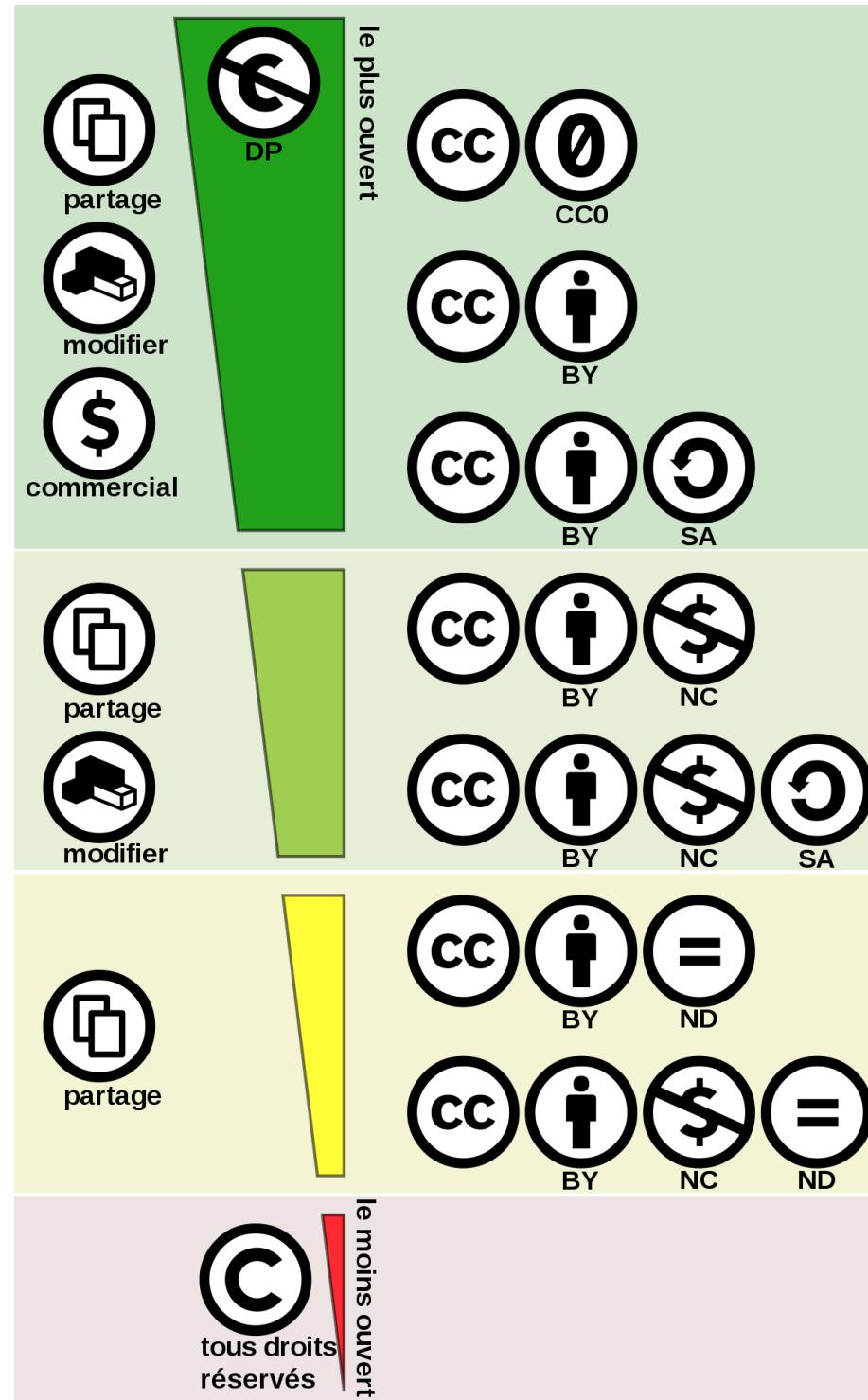
La [position officielle](#) de la Fondation Wikimedia est que « *les représentations fidèles des œuvres d'art du domaine public en deux dimensions sont dans le domaine public et les exigences contraires sont une attaque contre le concept même de domaine public* ». Pour plus de détails, voir

[Commons:Quand utiliser le bandeau PD-Art](#).

Cette reproduction photographique est donc également considérée comme étant élevée dans le domaine public.

Merci de noter qu'en fonction des lois locales, la réutilisation de ce contenu peut être interdite ou restreinte dans votre juridiction. Voyez

[Commons:Reuse of PD-Art photographs](#).



Diffuser les données

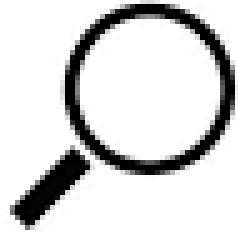
Réaliser techniquement un projet ne suffit pas: il faut préparer la distribution et la conservation des données, surtout dans une démarche *open*.

- Documenter: il faut expliquer comment s'organisent les données pour que n'importe qui (ou presque) puisse (idéalement) remonter le projet
- Distribuer: il existe différentes options comme le logiciel (payant?) et l'application en ligne (gratuite? mais qui paye?). Il faut faire attention aux langages de programmation utilisés (Python ou Angular?) et différencier la publication de l'application et du code de l'application avec les données
- Entreposer: il faut s'organiser pour que nos données nous survivent en utilisant des entrepôts sécurisés pour les données (type Zenodo) ou les publications (type HAL)

FAIR

F

indable



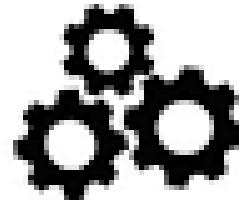
A

ccessible



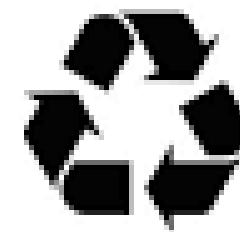
I

nteroperable



R

eusable



- F (*findable*) car les données doivent être facilement trouvables, par un permalien par exemple et des métadonnées pour les moteurs de recherche
- A (*Accessible*) car les données doivent être récupérables par un protocole de communication (IIIF, OAI-PMH...), assorties d'une licence claire, et avec des données accessibles si les données sont sous embargo
- I (*Interoperable*) car les données doivent suivre des standards connus
- R (*Reusable*) car les données doivent autant que possible être réutilisable, notamment grâce à une documentation claire, en plus des points précédemment évoqués (standards, métadonnées...)

Open

L'*open science* est un enjeu fondamental, tant d'un point de vue numérique que scientifique.

- *open access*: 
- *open source*: 
- *open data*: 
- *etc.*

L'intérêt de la démarche *open* est citoyen (gratuité) mais aussi scientifique (reproductibilité, garantie qualité).

Chaîne de traitement

Chaîne de traitement

On parle de chaîne de traitement (ou "flux de travail" selon la Commission générale de terminologie et de néologie) ou de *workflow*. Comme aucune solution informatique ne permet de tout faire (à l'inverse de logiciel comme *Word* en bureautique), il faut trouver

1. Une série de solutions...
2. ... qui s'articulent correctement les unes avec les autres...
3. ... et qui correspondent à des standards.

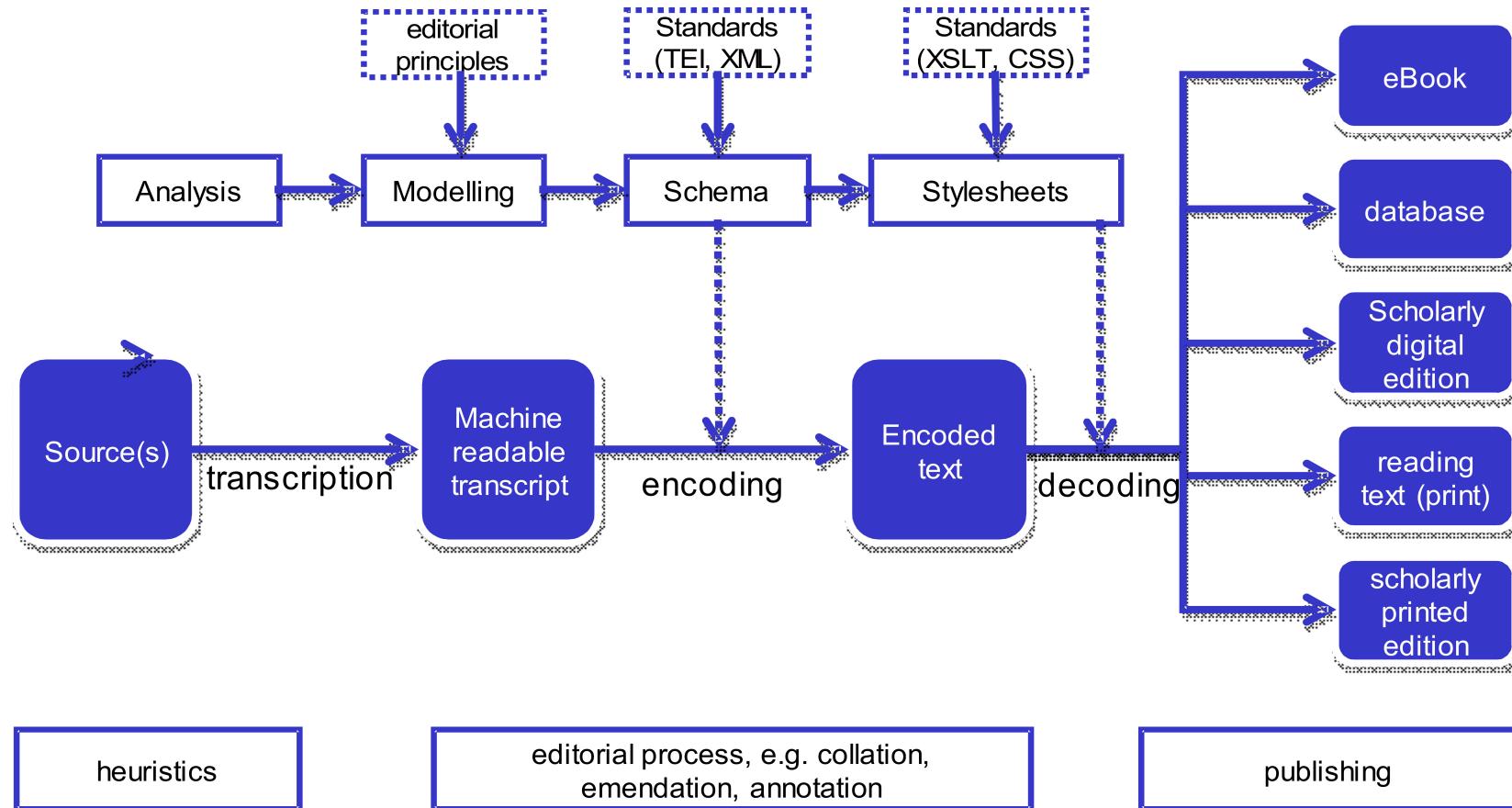
Pourquoi dois-je (presque) tout faire?

L'objectif d'avoir un équivalent de *Word* n'est pas nécessairement souhaitable. Toute simplification se paye:

- Au sens propre avec l'apparition de solutions privées, donc payantes.
- Au sens figuré, avec l'enfermement dans une solution générale qui gère mal les cas particuliers.

Cependant, il existe déjà des chaînes de traitement fonctionnelles et de très grande qualité, comme **METOPES** à l'université de Caen.

Un exemple de chaîne de traitement pour l'édition numérique



(Quelle: Rehbein/Fritze 2012)

Source: Christof Schöch, *Digitale Textedition mit TEI, en ligne*.

La place du numérique dans les discipline

L'édition numérique reprend les étapes de la philologie traditionnelle. Elle ouvre de nouvelles potentialités, malheureusement au prix d'une complexification du travail.

Retour à la renaissance, ou, comme Alde Manuce, l'humaniste maîtrise l'intégralité de la chaîne de production, de la transcription à la publication, en passant par la fabrication des outils (presse, fontes...).

L'édition numérique est avant tout une édition, et nécessite des compétences en ecdotique traditionnelle.

Quelques grandes étapes

1. Transcription -> Kraken, Ocropy, Tesseract...
2. Collation -> Collatex, Juxta...
3. Analyse paléographique -> Archetype...
4. Annotation linguistique -> TreeTagger, Marmot, Pie...
5. Exploitation linguistique -> TXM, Unitex/GramLab...
6. Exploitation littéraire -> Pour les emprunts: Tracer ou Philologic
7. Indexation -> HER, GROBID entity fishing...
8. Publication -> TEIPublisher, Synoptix, LaTeX
9. Archivage -> HAL, Huma-num

Bibliographie

- Huma-num, *Le guide de bonnes pratiques numériques*, version de 2015, https://www.huma-num.fr/sites/default/files/guide_des_bonnes_pratiques.pdf
- *Numériser et mettre en ligne*, sous la dir. de Thierry Claerr et Isabelle Westeel, Presses de l'Enssib, 2010.
<https://books.openedition.org/pressesenssib/414>